Genotypic variability enhances the reproducibility of an ecological study

Article

Accepted Version

Milcu, A., Puga-Freitas, R., Ellison, A. M., Blouin, M., Scheu, S., Freschet, G. T., Rose, L., Barot, S., Cesarz, S., Eisenhauer, N., Girin, T., Assandri, D., Bonkowski, M., Buchmann, N., Butenschoen, O., Devidal, S., Gleoxner, G., Gessler, A., Gigon, A., Greiner, A., Grignani, C., Hansart, A., Kayler, Z., Lange, M., Lata, J. C., Le Galliard, J. F., Lukac, M. ORCID: https://orcid.org/0000-0002-8535-6334, Mannerheim, N., Muller, M. E. H., Pando, A., Rotter, P., Scherer-Lorenzen, M., Seyhun, R., Urban-Maed, K., Weigelt, A., Zavattaro, L. and Roy, J. (2018) Genotypic variability enhances the reproducibility of an ecological study. Nature Ecology & Evolution, 2 (2). pp. 279-287. ISSN 2397-334X doi: https://doi.org/10.1038/s41559-017-0434-x Available at https://centaur.reading.ac.uk/74258/

Genotypic variability enhances the reproducibility of an ecological study

Alexandru Milcu[1,2], Ruben Puga-Freitas[3], Aaron M. Ellison[4,5], Manuel Blouin[3,6], Stefan Scheu[7], Grégoire T. Freschet[2], Laura Rose[8], Sebastien Barot[9], Simone Cesarz[10,11], Nico Eisenhauer[10,11], Thomas Girin[12], Davide Assandri[13], Michael Bonkowski[14], Nina Buchmann[15], Olaf Butenschoen[7,16], Sebastien Devidal[1], Gerd Gleixner[17], Arthur Gessler[18,19], Agnès Gigon[3], Anna Greiner[8], Carlo Grignani[13], Amandine Hansart[20], Zachary Kayler[19,21], Markus Lange[17], Jean-Christophe Lata[22], Jean-François Le Galliard[20,22], Martin Lukac[23,24], Neringa Mannerheim[15], Marina E.H. Müller[18], Anne Pando[6], Paula Rotter[8], Michael Scherer-Lorenzen[8], Rahme Seyhun[22], Katherine Urban-Mead[2], Alexandra Weigelt[10,11], Laura Zavattaro[13] and Jacques Roy[1]

[1]Ecotron (UPS-3248), CNRS, Campus Baillarguet, F-34980, Montferrier-sur-Lez, France.
[2]Centre d'Ecologie Fonctionnelle et Evolutive, CEFE-CNRS, UMR 5175, Université de Montpellier – Université Paul Valéry – EPHE, 1919 route de Mende, F-34293, Montpellier Cedex 5, France.
[3]Institut des Sciences de l'Ecologie et de l'Environnement de Paris (UPMC, UPEC, Paris Diderot, CNRS, IRD, INRA), Université Paris-Est Créteil, 61 avenue du Général De Gaulle, F-94010 Créteil Cedex, France.
[4]Harvard Forest, Harvard University, 324 North Main Street, Petersham, Massachusetts, USA.
[5]University of the Sunshine Coast, Tropical Forests and People Research Centre, Locked Bag 4, Maroochydore DC, Queensland 4558, Australia.
[6]Agroécologie, AgroSup Dijon, INRA, Univ. Bourgogne Franche-Comté, F-21000 Dijon, France
[7]J.F. Blumenbach Institute for Zoology and Anthropology, Georg August University Göttingen, Berliner Str. 28, 37073 Göttingen, Germany.
[8]Faculty of Biology, University of Freiburg, Geobotany, Schaenzlestr. 1, D-79104 Freiburg, Germany.
[9]IRD, Institut des Sciences de l'Ecologie et de l'Environnement de Paris (UPMC, UPEC, Paris Diderot, CNRS, IRD, INRA), UPMC, Bâtiment 44-45, deuxième étage, bureau 208, CC 237, 4 place Jussieu, 75252 Paris cedex 05, France.

[10]German Centre for Integrative Biodiversity Research (iDiv), Halle-Jena-Leipzig, Deutscher Platz 5e, 04103 Leipzig, Germany.
[11]Institute of Biology, Leipzig University, Deutscher Platz 5e, 04103 Leipzig, Germany.
[12]Institut Jean-Pierre Bourgin, INRA, AgroParisTech, CNRS, Université Paris-Saclay, RD10, 78026 Versailles Cedex, France.
[13]Department of Agricultural, Forest and Food Sciences, University of Turin, largo Braccini, 2, 10095 Grugliasco, Italy.
[14]Cluster of Excellence on Plant Sciences (CEPLAS), Terrestrial Ecology Group, Institute for Zoology, University of Cologne, Zülpicher Str. 47b, 50674 Köln, Germany.
[15]Institute of Agricultural Sciences, ETH Zurich, Universitätsstrasse 2, 8092 Zürich, Switzerland
[16]Senckenberg Biodiversität und Klima Forschungszentrum BiK-F, Georg-Voigt-Straße 14-16, Frankfurt am Main.
[17]Max Planck Institute for Biogeochemistry, Postfach 100164, 07701 Jena, Germany.
[18]Leibniz Centre for Agricultural Landscape Research (ZALF), Institute of Landscape Biogeochemistry, Eberswalder Str. 84, 15374 Müncheberg, Germany.
[19]Swiss Federal Research Institute WSL, Zürcherstr. 111, 8903 Birmensdorf, Switzerland.
[20]Ecole normale supérieure, PSL Research University, Département de biologie, CNRS, UMS 3194, Centre de recherche en écologie expérimentale et prédictive (CEREEP-Ecotron IleDeFrance), 78 rue du château, 77140 Saint-Pierre-lès-Nemours, France.
[21]Department of Soil and Water Systems, University of Idaho, 875 Perimeter Dr., Moscow, ID, USA.
[22]Institut des Sciences de l'Ecologie et de l'Environnement de Paris (UPMC, UPEC, Paris Diderot, CNRS, IRD, INRA), Sorbonne Universités, CC 237, 4 place Jussieu, 75252 Paris cedex 05, France.
[23]School of Agriculture, Policy and Development, University of Reading, Reading, RG6 6AR, UK.
[24]FLD, Czech University of Life Sciences, 165 00 Prague, Czech Republic.

Corresponding author: Alexandru Milcu, CNRS, Ecotron - UPS 3248, Campus Baillarguet, 34980, Montferrier-sur-Lez, France, email: alex.milcu@cnrs.fr, phone: +33 (0) 434-359-893.

Many scientific disciplines are currently experiencing a "reproducibility crisis" because numerous scientific findings cannot be repeated consistently. A novel but controversial hypothesis postulates that stringent levels of environmental and biotic standardization in experimental studies reduces reproducibility by amplifying impacts of lab-specific environmental factors not accounted for in study designs. A corollary to this hypothesis is that a deliberate introduction of controlled systematic variability (CSV) in experimental designs may lead to increased reproducibility. We tested this hypothesis using a multi-laboratory microcosm study in which the same ecological experiment was repeated in 14 laboratories across Europe. Each laboratory introduced environmental and genotypic CSV within and among replicated microcosms established in either growth chambers (with stringent control of environmental conditions) or glasshouses (with more variable environmental conditions). The introduction of genotypic CSV led to lower among-laboratory variability in growth chambers, indicating increased reproducibility, but had no significant effect in glasshouses where reproducibility was generally lower. Environmental CSV had little effect on reproducibility. Although there are multiple causes for the "reproducibility crisis", deliberately including genetic variation may be a simple solution for increasing the reproducibility of ecological studies performed in controlled environments.

Reproducibility—the ability to duplicate a study and its findings—is a defining feature of scientific research. In ecology, it is often argued that it is virtually impossible to accurately duplicate any single ecological experiment or observational study. The rationale is that the complex ecological interactions between the ever-changing environment and the extraordinary

diversity of biological systems exhibiting a wide range of plastic responses at different levels of biological organization make exact duplication unfeasible[1,2]. Although this may be true for observational and field studies, numerous ecological (and agronomic) studies are carried out with artificially assembled simplified ecosystems and controlled environmental conditions in experimental microcosms or mesocosms (henceforth, "microcosms")[3–5]. Since biotic and environmental parameters can be tightly controlled in microcosms, results from such studies should be easier to reproduce. Even though microcosms have frequently been used to address fundamental ecological questions[4,6,7], there has been no quantitative assessment of the reproducibility of any microcosm experiment.

Experimental standardization— the implementation of strictly defined and controlled properties of organisms and their environment—is widely thought to increase both reproducibility and sensitivity of statistical tests[8,9] because it reduces within-treatment variability. This paradigm has been recently challenged by several studies on animal behavior, suggesting that stringent standardization may, counterintuitively, be responsible for generating non-reproducible results[9–11] and contribute to the actual reproducibility crisis[12–15]; the results may be valid under given conditions (i.e., they are local "truths") but are not generalizable[8,16]. Despite rigorous adherence to experimental protocols, laboratories inherently vary in many conditions that are not measured and are thus unaccounted for, such as experimenter, micro-scale environmental heterogeneity, physico-chemical properties of reagents and lab-ware, pre-experimental conditioning of organisms, and their genetic and epigenetic background. It even has been suggested that attempts to stringently control all sources of biological and environmental variation might inadvertently lead to the amplification of the effects of these unmeasured variations among laboratories, thus reducing reproducibility[9–11].

104 Some studies have gone even further, hypothesizing that the introduction of controlled
105 systematic variation (CSV) among the replicates of a treatment (e.g., using different genotypes or
106 varying the organisms' pre-experimental conditions among the experimental replicates) should
107 lead to less variable mean response values between the laboratories that duplicate the
108 experiments[9,11]. In short, it has been argued that reproducibility may be improved by shifting the
109 variance from among experiments to within them[9]. If true, then introducing CSV will increase
110 researchers' ability to draw generalizable conclusions about the directions and effect sizes of
111 experimental treatments and reduce the probability of false positives. The trade-off inherent to
112 this approach is that increasing within-experiment variability will reduce the sensitivity (i.e. the
113 probability of detecting true positives) of statistical tests. However, it currently remains unclear
114 whether introducing CSV increases reproducibility of ecological microcosm experiments, and if
115 so, at what cost for the sensitivity of statistical tests.

116 To test the hypothesis that introducing CSV enhances reproducibility in an ecological
117 context, we had 14 European laboratories simultaneously run a simple microcosm experiment
118 using grass (*Brachypodium distachyon* L.) monocultures and grass and legume (*Medicago*
119 *truncatula* Gaertn.) mixtures. As part of the reproducibility experiment, the 14 laboratories
120 independently tested the hypothesis that the presence of the legume species *M. truncatula* in
121 mixtures would lead to higher total plant productivity in the microcosms and enhanced growth of
122 the non-legume *B. distachyon via* rhizobia-mediated nitrogen fertilization and/or nitrogen
123 sparing effects[17–19].

124 All laboratories were provided with the same experimental protocol, seed stock from the
125 same batch, and identical containers in which to establish microcosms with grass only and grass-
126 legume mixtures. Alongside a control (CTR) with no CSV and containing a homogenized soil

127 substrate (mixture of soil and sand) and a single genotype of each plant species, we explored the
128 effects of five different types of within- and among-microcosm CSV on experimental
129 reproducibility of the legume effect (Fig. 1): 1) within-microcosm environmental CSV (ENV_W)
130 achieved by spatially varying soil resource distribution through the introduction of six sand
131 patches into the soil; 2) among-microcosm environmental CSV (ENV_A), which varied the
132 number of sand patches (none, three, or six) among replicate microcosms; 3) within-microcosm
133 genotypic CSV (GEN_W) that used three distinct genotypes per species planted in homogenized
134 soil in each microcosm; 4) among-microcosm genotypic CSV (GEN_A) that varied the number of
135 genotypes (one, two, or three) planted in homogenized soil among replicate microcosms; and 5)
136 both genotypic and environmental CSV (GEN_W+ENV_W) within microcosms that used six sand
137 patches and three plant genotypes per species in each microcosm. In addition, we tested whether
138 CSV effects are modified by the level of standardization within laboratories by using two
139 common experimental approaches ('SETUP' hereafter): growth chambers with tightly controlled
140 environmental conditions and identical soil (eight laboratories) or glasshouses with more loosely
141 controlled environmental conditions and different soils (six laboratories; see Supplementary
142 Table 1 for the physico-chemical properties of the soils).

143 We measured 12 parameters representing a typical ensemble of response variables reported
144 for plant-soil microcosm experiments. Six of these were measured at the microcosm-level: shoot
145 biomass, root biomass, total biomass, shoot-to-root ratio, evapotranspiration, and decomposition
146 of a common substrate using a simplified version of the "teabag litter decomposition method"[20].
147 The other six were measured on *B. distachyon* alone: seed biomass, height, and four shoot-tissue
148 chemical variables; N%, C%, $\delta^{15}N$, $\delta^{13}C$. All 12 variables were then used to calculate the effect
149 of the presence of a nitrogen-fixing legume on ecosystem functions in grass-legume mixtures

150 ('net legume effect' hereafter) (Supplementary Table 2), calculated as the difference between the
151 values measured in the microcosms with and without legumes, an approach often used in
152 legume-grass binary cropping systems[19,21] and biodiversity-ecosystem function experiments[17,22].
153 Statistically significant differences among the 14 laboratories were considered an indication
154 of irreproducibility. In the first instance, we assessed how our experimental treatments (CSV and
155 SETUP) affected the number of laboratories that produced results that could be considered to
156 have reproduced the same finding. We then determined how experimental treatments affected
157 standard deviation (SD) of the legume effect for each of the 12 variables both within- and
158 among-laboratories; lower among-laboratory SD implies that the results were more similar,
159 suggesting increased reproducibility. Lastly, we explored the relationship between within- and
160 among-laboratory SD, and how the experimental treatments affected the statistical power of
161 detecting the net legume effect.
162 Although each laboratory followed the same experimental protocol, we found a remarkably
163 high level of among-laboratory variation for most response variables (Supplementary Fig. 1) and
164 the net legume effect on those variables (Fig. 2). For example, the net legume effect on mean
165 total plant biomass varied among laboratories from 1.31 to 6.72 g dry weight (DW) per
166 microcosm in growth chambers, suggesting that unmeasured laboratory-specific conditions
167 outweighed effects of experimental standardization. Among glasshouses, differences were even
168 larger: the net legume effect on mean plant biomass varied by two orders of magnitude, from
169 0.14 to 14.57g DW per microcosm (Fig. 2). Furthermore, for half of the variables (root biomass,
170 litter decomposition, grass height, foliar C%, $\delta^{15}C$ and $\delta^{15}N$) the direction of the net legume
171 effect varied with laboratory.

Mixed-effects models were used to test the effect of legume species presence (LEG), laboratory (LAB), CSV, and their interactions (with experimental block—within-LAB growth chamber or glasshouse bench—as a random factor) on the 12 response variables. The impact of the presence of legumes varied significantly with laboratory and CSV for half of the variables, as indicated by the LEG×LAB×CSV three-way interaction (Table 1, Supplementary Figs 2 and 3). For the other half, significant two-way interactions between LEG×LAB and CSV×LAB were found. The same significant interactions were found when analyzing the first (PC1) and second (PC2) principal components from a principal component analysis (PCA) that included all 12 response variables; PC1 and PC2 together explained 45% of the variation (Table 1; Supplementary Fig. 4ab). Taken together, these results suggest that the effect size or direction of the net legume effect was significantly different (i.e. not reproducible) in some laboratories and that the introduced CSV treatment affected reproducibility. In a complementary analysis including the SETUP in the model (and accounting for the LAB effect as a random factor), we found that the impact of the CSV treatment varied significantly with the SETUP (CSV×SETUP or LEG×CSV×SETUP interactions; Supplementary Table 3), suggesting the reproducibility of the results differed between glasshouses and growth chambers.

To answer the question of how many laboratories produced results that were statistically indistinguishable from one another (i.e. reproduced the same finding), we used Tukey's post-hoc Honest Significant Difference (HSD) test for the LAB effect on the first and second principal components describing the net legume effect, which together explained 49% of the variation (Supplementary Fig. 4cd). Out of the 14 laboratories, seven (PC1) and 11 (PC2) laboratories were statistically indistinguishable in controls; this value increased in the treatments with environmental or genotypic CSV for PC1 but not PC2 (Table 2). When we analyzed responses in

growth chambers alone, five of eight laboratories were statistically indistinguishable in controls, but this increased to six out of eight laboratories when we considered treatments with only environmental CSV and seven of eight in treatments with genotypic CSV (GEN_W, GEN_A and GEN_W+ENV_W). In glasshouses, introducing CSV did not affect the number of statistically indistinguishable laboratories with respect to PC1 but decreased the number of statistically indistinguishable laboratories with respect to PC2 (Table 2).

We also assessed the impact of the experimental treatments on the among- and within-laboratory SD. Analysis of the among-laboratory SD of the net legume effect revealed a significant CSV×SETUP interaction ($F_{5,121}$=7.38, P < 0.001) (Fig. 3a, b). This interaction included significantly lower fitted coefficients (i.e., lower among-laboratory SD) in growth chambers for GEN_W ($t_{5,121}$ = -3.37, P = 0.001), GEN_A ($t_{5,121}$ = -2.95, P = 0.004) and ENV_W+GEN_W ($t_{1,121}$ = -3.73, P < 0.001) treatments relative to CTR (see full model output for among-laboratory SD in Supplementary Note). For these three treatments, the among-laboratory SD of the net legume effect was 18% lower with genotypic CSV than without it, indicating increased reproducibility (Fig. 3a). The same analysis performed on within-laboratory SD of the net legume effect only found a slight but significant increase of within-laboratory SD in the GEN_A treatment ($t_{5,121}$ = 3.52, P < 0.001) (see model output for within-laboratory SD in Supplementary Note). We then tested whether there was a relationship between within- and among-laboratory SD with a statistical model for among-laboratory SD as a function of within-laboratory SD, SETUP, CSV and their interactions. We found a significant within-laboratory SD×SETUP×CSV three-way interaction ($F_{5,109}$ = 2.4, P < 0.040) affecting among-laboratory SD (Supplementary Note). This interaction was the result of a more negative relationship between

217 within- and among-laboratory SD in glasshouses relative to growth chambers, but with different
218 slopes for the different CSV treatments (Fig. 4).

219 Introducing CSV can increase within-laboratory variation, as indicated by the positive
220 coefficients fitted in some of the CSV treatments in the model output for within-laboratory SD
221 (see Supplementary Note). Thus, for the three CSV treatments that produced the most consistent
222 results (GEN_W, GEN_A, ENV_W+GEN_W), we analyzed the statistical power of detecting the net
223 legume effect within individual laboratories. In growth chambers, adding genotypic CSV led to a
224 slight reduction in statistical power relative to CTR (57% in CTR vs. 46% in the three treatments
225 containing genotypic variability) that could have been compensated for by using eleven instead
226 of six replicated microcosms per treatment. In glasshouses, owing to a higher effect size of
227 legume presence on the response variables, the statistical power for detecting the legume effect
228 in CTR was slightly higher (68%) than in growth chambers, but was reduced to 51% on average
229 for the three treatments containing genotypic CSV, a decrease that could have been compensated
230 for by using 16 replicated microcosms instead of six.

231 Overall, our study shows that results produced by microcosm experiments can be strongly
232 biased by lab-specific factors. Based on the principal component explaining most of the variation
233 in the twelve response variables (PC1), only seven out of the 14 laboratories produced results
234 that can be considered reproducible (Table 2) with the current standardization procedures. This
235 result is in line with the only other comparable study[12] (to the best of our knowledge) reporting
236 that out of ten laboratories, only four generated similar leaf growth phenotypes of *Arabidopsis*
237 *thaliana* (L). In addition to highlighting that approximately one in two ecological studies
238 performed in microcosms under controlled environments produce statistically different results,
239 our study provides supporting evidence for the hypothesis that introducing genotypic CSV can

increase reproducibility of ecological studies[9–11]. However, the effectiveness of genotypic CSV for enhancing reproducibility varied with the setup; it led to lower (−18%) among-laboratory SD in growth chambers only, with no benefit observed in glasshouses. Lower among-laboratory SD in growth chambers implies that the microcosms containing genotypic CSV were less strongly affected by unaccounted-for lab-specific environmental or biotic variables. Analyses performed at the level of individual variables (Table 1) showed that introducing genotypic CSV affected the among-laboratory SD in most, but not all variables. This suggests that the relationship between genotypic CSV and reproducibility is probabilistic and results from the decreased likelihood that microcosms containing CSV will respond to unaccounted for lab-specific environmental factors in the same direction and with the same magnitude. The mechanism is likely to be analogous to the stabilizing effect of biodiversity on ecosystem functions under changing environmental conditions[23–26], but additional empirical evidence is needed to confirm this conjecture.

Introducing genotypic CSV increased reproducibility in growth chambers (with stringent control of environmental conditions) but not in glasshouses (with more variable environmental conditions). Higher among-laboratory SD in glasshouses may indicate the existence therein of stronger laboratory-specific factors, and our deliberate use of different soils in the glasshouses presumably contributed to this effect. However, the among-laboratory SD in glasshouses decreased with increasing within-laboratory SD, irrespective of CSV, an effect that was less clear in growth chambers (Fig. 4). This observation appears to be in line with the hypothesis put forward by Richter et al.[9], who proposed that increasing the variance within experiments can reduce the among-laboratory variability of the mean effect sizes observed in each laboratory. Yet, despite the negative correlation between within- and among-laboratory SD observed in glasshouses, the among-laboratory SD remained higher in glasshouses than in growth chambers.

263 Therefore, we consider that the hypothesized mechanistic link between CSV-induced higher
264 within-laboratory SD and increased reproducibility is poorly supported by our dataset.
265 Nevertheless, one possible explanation for the lack of effect on reproducibility in glasshouses is
266 that our CSV treatments did not introduce a sufficiently high level of within-laboratory
267 variability to buffer against laboratory-specific factors for all response variables; across the
268 twelve response variables, the average main effect (i.e., without the interaction terms) of the
269 CSV treatment contributed to a low percentage (2.6% ± 1.6 s.e.m.) of the total sum of squares
270 relative to the main effects of laboratory (43.4% ± 5.2 s.e.m.) and legumes (10.9% ± 3.1 s.e.m.).
271 A similar conjecture was put forward by the other two studies that explored the role of CSV for
272 reproducibility in animal behavior[9,10]. At present we are unable to conclude that the introduction
273 of stronger sources of controlled within-laboratory variability can increase reproducibility in
274 glasshouses with more loosely controlled environmental conditions and different soils.
275 Our results indicate that genotypic CSV is more effective in increasing reproducibility than
276 environmental CSV, irrespective of whether the CSV was introduced within or among individual
277 replicates (i.e., microcosms). However, we cannot discount the possibility that we found this
278 result because our treatments with environmental CSV were less successful in increasing within-
279 microcosm variability. Additional experiments could test whether other types of environmental
280 CSV, such as soil nutrients, texture, or water availability, might be more effective at increasing
281 reproducibility.
282 We expected higher overall productivity (i.e., a net legume effect) in the grass-legume
283 mixtures and enhanced growth of *B. distachyon* because of the presence of the nitrogen (N)-
284 fixing *M. truncatula*. However, these species were not selected because of their routine pairings
285 in agronomic or ecological experiments (they are rarely used that way), but rather because they

286 are frequently present in controlled environment experiments looking at functional genomics.
287 Contrary to our expectation, and despite the generally lower ^{15}N signature of *B. distachyon* in the
288 presence of N-fixing *M. truncatula* (suggesting that some of the N fixed by *M. truncatula* was
289 taken up by the grass), the biomass of *B. distachyon* was lower in the microcosms containing *M.*
290 *truncatula*. Seed mass and shoot %N data of *B. distachyon* was lower in mixtures
291 (Supplementary Fig. 1), suggesting that the two species competed for N. The lack of a significant
292 N fertilization effect of *M. truncatula* on *B. distachyon* could have resulted from the
293 asynchronous phenologies of the two species: the 8–10-week life cycle of *B. distachyon* may
294 have been too short to benefit from the N fixation by *M. truncatula*.
295 Because well-established meta-analytical approaches can account for variation caused by
296 local factors and still detect the general trends across different types of experimental setups,
297 environments, and populations, we should ask whether the additional effort required for
298 introducing CSV in experiments is worthwhile. Considering the current reproducibility crisis in
299 many fields of science[27], we suggest that it is, for at least three reasons. First, some studies
300 become seminal without any attempts to reproduce them. Second, even if a seminal study that is
301 flawed due to laboratory-specific biases is later proven wrong, it usually takes significant time
302 and resources before its impact on the field abates. Third, the current rate of reproducibility is
303 estimated to be as low as one-third[12–14], implying that most data entering any meta-analysis are
304 biased by unknown lab-specific factors. Addition of genotypic CSV may enhance the
305 reproducibility of individual experiments and eliminate potential biases in data used in meta-
306 analyses. Last, if each individual study is less affected by laboratory-specific unknown
307 environmental and biotic factors, then we would also need fewer studies to draw solid
308 conclusions about the generality of phenomena. Therefore, we argue that investing more in

making individual studies more reproducible and generalizable will be beneficial in both the short and long run. At the same time, adding CSV can reduce statistical power to detect experimental effects, so some additional experimental replicates would be needed when using it.

Arguably, our use of statistical significance tests of effects sizes to determine reproducibility might be viewed as overly restrictive and better suited to assessing reproducibility of parameter estimates rather than assessing the generality of the hypothesis under test[27]. We used this approach because no generally accepted alternative framework is available to assess how close the multivariate results from multiple laboratories need to be to conclude that they reproduced the same finding. It is worth noting that although the direction of the legume effect was the same in the majority of laboratories, the differences among laboratories were very large (e.g., up to two orders of magnitude for shoot biomass) and in 10% of the 168 laboratory × variable combinations (14 laboratories × 12 response variables) the direction of the legume effect differed from the among-laboratory consensus (Fig. 2).

In conclusion, our study shows that the current standardization procedures used in ecological microcosm experiments are inadequate in accounting for lab-specific environmental factors and suggests that introducing controlled variability in experiments may buffer effects of lab-specific factors. Although there are multiple causes for the reproducibility crisis[15,28,29], deliberately including genetic variation in the studied organisms can be a simple solution for increasing the reproducibility of ecological studies performed in controlled environments. However, as the introduced genotypic variability only increased reproducibility in experimental setups with tightly controlled environmental conditions (i.e., in growth chambers using identical soil), our study indicates that the reproducibility of ecological experiments can be enhanced by a

combination of rigorous standardization of environmental variables at the laboratory level as well as controlled genotypic variability.

References

1. Cassey, P. & Blackburn, T. Reproducibility and Repeatability in Ecology. *Bioscience* **56,** 958–9 (2006).
2. Ellison, A. M. Repeatability and transparency in ecological research. *Ecology* **91,** 2536–2539 (2010).
3. Lawton, J. H. The Ecotron facility at Silwood Park: the value of‘ big bottle’ experiments. *Ecology* **77,** 665–669 (1996).
4. Benton, T. G., Solan, M., Travis, J. M. & Sait, S. M. Microcosm experiments can inform global ecological problems. *Trends Ecol. Evol.* **22,** 516–521 (2007).
5. Drake, J. M. & Kramer, A. M. Mechanistic analogy: how microcosms explain nature. *Theor. Ecol.* **5,** 433–444 (2012).
6. Fraser, L. H. & Keddy, P. The role of experimental microcosms in ecological research. *Trends Ecol. Evol.* **12,** 478–481 (1997).
7. Srivastava, D. S. *et al.* Are natural microcosms useful model systems for ecology? *Trends Ecol. Evol.* **19,** 379–384 (2004).
8. De Boeck, H. J. *et al.* Global change experiments: challenges and opportunities. *Bioscience* (2015). doi:10.1093/biosci/biv099
9. Richter, S. H. *et al.* Effect of population heterogenization on the reproducibility of mouse behavior: a multi-laboratory study. *PLoS One* **6,** e16461 (2011).
10. Richter, S. H., Garner, J. P. & Würbel, H. Environmental standardization: cure or cause of

poor reproducibility in animal experiments? *Nat. Methods* **6,** 257–261 (2009).

11. Richter, S. H., Garner, J. P., Auer, C., Kunert, J. & Würbel, H. Systematic variation improves reproducibility of animal experiments. *Nat. Methods* **7,** 167–8 (2010).

12. Massonnet, C. *et al.* Probing the reproducibility of leaf growth and molecular phenotypes: a comparison of three Arabidopsis accessions cultivated in ten laboratories. *Plant Physiol.* **152,** 2142–2157 (2010).

13. Begley, C. G. & Ellis, M. L. Raise standards for preclinical cancer research. *Nature* **483,** 531–533 (2012).

14. Open Science Collaboration. Estimating the reproducibility of psychological science. *Science (80-.).* **349,** aac4716 (2015).

15. Parker, T. H. *et al.* Transparency in ecology and evolution: real problems, real solutions. *Trends Ecol. Evol.* **31,** 711–719 (2016).

16. Moore, R. P. & Robinson, W. D. Artificial bird nests, external validity, and bias in ecological field studies. *Ecology* **85,** 1562–1567 (2004).

17. Temperton, V. M., Mwangi, P. N., Scherer-Lorenzen, M., Schmid, B. & Buchmann, N. Positive interactions between nitrogen-fixing legumes and four different neighbouring species in a biodiversity experiment. *Oecologia* **151,** 190–205 (2007).

18. Meng, L. *et al.* Arbuscular mycorrhizal fungi and rhizobium facilitate nitrogen uptake and transfer in soybean/maize intercropping system. *Front. Plant Sci.* **6,** 339 (2015).

19. Sleugh, B., Moore, K. J., George, J. R. & Brummer, E. C. Binary legume–grass mixtures improve forage yield, quality, and seasonal distribution. *Agron. J.* **92,** 24–29 (2000).

20. Keuskamp, J. a., Dingemans, B. J. J., Lehtinen, T., Sarneel, J. M. & Hefting, M. M. Tea Bag Index: a novel approach to collect uniform decomposition data across ecosystems.

Methods Ecol. Evol. **4,** 1070–1075 (2013).

21. Nyfeler, D., Huguenin-Elie, O., Suter, M., Frossard, E. & Lüscher, A. Grass-legume mixtures can yield more nitrogen than legume pure stands due to mutual stimulation of nitrogen uptake from symbiotic and non-symbiotic sources. *Agric. Ecosyst. Environ.* **140,** 155–163 (2011).

22. Suter, M. *et al.* Nitrogen yield advantage from grass-legume mixtures is robust over a wide range of legume proportions and environmental conditions. *Glob. Chang. Biol.* **21,** 2424–2438 (2015).

23. Loreau, M. & de Mazancourt, C. Biodiversity and ecosystem stability: A synthesis of underlying mechanisms. *Ecol. Lett.* **16,** 106–115 (2013).

24. Reusch, T. B., Ehlers, A., Hämmerli, A. & Worm, B. Ecosystem recovery after climatic extremes enhanced by genotypic diversity. *Proc. Natl. Acad. Sci. U. S. A.* **102,** 2826 (2005).

25. Hughes, A. R., Inouye, B. D., Johnson, M. T. J., Underwood, N. & Vellend, M. Ecological consequences of genetic diversity. *Ecol. Lett.* **11,** 609–623 (2008).

26. Prieto, I. *et al.* Complementary effects of species and genetic diversity on productivity and stability of sown grasslands. *Nat. Plants* **1,** 1–5 (2015).

27. Wasserstein, R. L. & Lazar, N. A. The ASA's statement on p-values: context, process, and purpose. Am. Stat. 70, 129–133 (2016).

28. Baker, M. 1,500 scientists lift the lid on reproducibility. *Nature* **533,** 452–454 (2016).

29. Nuzzo, R. How scientists fool themselves – and how they can stop. *Nature* **526,** 182–185 (2015).

Acknowledgements

This study benefited from the CNRS human and technical resources allocated to the ECOTRONS Research Infrastructures and the state allocation 'Investissement d'Avenir' ANR-11-INBS-0001 and from financial support by the ExpeER (grant no. 262060) consortium funded under the EU-FP7 research program (FP2007-2013). *Brachypodium* seeds were kindly provided by Richard Sibout (Observatoire du Végétal, Institut Jean-Pierre Bourgin, F-78026 Versailles Cedex France) and *Medicago* seeds were supplied by Jean-Marie Prosperi (INRA Biological Resource Centre, F-34060 Montpellier Cedex 1, France). We further thank Jean Varale, Gesa Hoffmann, Paul Werthenbach, Oliver Ravel, Clement Piel and Damien Landais, David Degueldre, Thierry Mathieu, Pierrick Aury, Nicolas Barthès, Bruno Buatois, Raphaëlle Leclerc for assistance duing the study. For additional acknowledgements see Supplementary Information.

Author contributions

A.M. and J.R. designed the study with input from M.B, S.B and J-C.L. Substantial methodological contributions were provided by M.B., S.S., T.G., L.R. and M.S-L. Conceptual feedback on an early version was provided by G.F., N.E., J.R. and A.M.E. Data were analysed by A.M. with input from A.M.E. A.M. wrote the manuscript with input from all co-authors. All co-authors were involved in carrying out the experiments and/or analyses.

Author Information

The authors declare no conflict of interest. Correspondence and request for materials should be addressed to Alexandru Milcu (alex.milcu@cnrs.fr).

METHODS

All laboratories tried to the best of their abilities to carry out an identical experimental protocol. Whereas not all laboratories managed to recreate precisely all details of the experimental protocol, we considered this to be a realistic scenario under which ecological experiments using microcosms are performed in glasshouses and growth chambers.

Germination

The seeds from the three genotypes of *Brachypodium distachyon* (Bd21, Bd21-3 and Bd3-1) and *Medicago truncatula* (L000738, L000530 and L000174) were first sterilized by soaking 100 seeds in 100 mL of a sodium hypochlorite solution with 2.6% active chlorine, and stirred for 15 min using a magnet. Thereafter, the seeds were rinsed 3 times in 250 mL of sterile water for 10-20 seconds under shaking. Sterilized seeds were germinated in trays (10 cm deep) filled with vermiculite. The trays were kept at 4°C in the dark for three days before being moved to light conditions (300 μmol m^{-2} s^{-1} PAR) and 20/16°C and 60/70% air RH for day- and night-time, respectively. When the seedlings of both species reached 1 cm in height above the vermiculite, they were transplanted into the microcosms.

Preparation of microcosms

All laboratories used identical containers (2-liter volume, 14.8-cm diameter, 17.4-cm height). Sand patches were created using custom-made identical “patch makers” consisting of six rigid PVC tubes (2.5 cm in diameter and 25 cm long), arranged in a circular pattern with an outer diameter of 10 cm. A textile mesh was placed at the bottom of the containers to prevent the spilling of soil through drainage holes. Filling of microcosms containing sand patches started with the insertion of the empty tubes into the containers. Thereafter, in growth chambers, 2000-g dry-weight of soil, subtracting the weight of the sand patches, was added into the containers and

around the “patch maker” tubes. Because different soils were used in the glasshouses, the dry weight of the soil differed depending on the soil density and was first estimated individually in each laboratory as the amount of soil needed to fill the pots up to 2 cm from the top. After the soil was added to the containers, the tubes were filled with a mixture of 10% soil and 90% sand. When the microcosms did not contain sand patches, the amount of sand otherwise contained in the six patches was homogenized with the soil. During the filling of the microcosms, a common substrate for measuring litter decomposition was inserted at the center of the microcosm at 8 cm depth. For simplicity as well as for its fast decomposition rate, we used a single batch of commercially available tetrahedron-shaped synthetic tea bags (mesh size of 0.25 mm) containing 2 g of green tea (Lipton, Unilever), as proposed by the “tea bag index” method[20]. Once filled, the microcosms were watered until water could be seen pouring out of the pot. The seedlings were then manually transplanted to predetermined positions (Fig. 1), depending on the genotype and treatment. Each laboratory established two blocks of 36 microcosms each, resulting in a total of 72 microcosms per laboratory, with blocks representing two distinct chambers in growth chamber setups or two distinct growth benches in the same glasshouse**.**

Soils

All laboratories using growth chamber setups used the same soil, whereas the laboratories using glasshouses used different soils (see Supplementary Table 1 for the physicochemical properties of the soils). The soil used in growth chambers was classified as a nutrient-poor cambisol and was collected from the top layer (0–20 cm) of a natural meadow at the Centre de Recherche en Ecologie Expérimentale et Prédictive—CEREEP (Saint-Pierre-Lès-Nemours, France). Soils used in glasshouses originated from different locations. The soil used by laboratory L2 was a fluvisol collected from the top layer (0-40 cm) of a quarry site near Avignon, in the Rhône valley,

Southern France. The soil used by laboratory L4 was collected from near the La Cage field experimental system (Versailles, France) and was classified as a luvisol. The soil used by labs L11 and L12 was collected from the top layer (0-20cm) within the haugh of the river Dreisam in the East of Freiburg, Germany. This soil was classified as an umbric gleysol with high organic carbon content. The soil from laboratory L14 was classified as a eutric fluvisol and was collected on the field site of the Jena Experiment, Germany. Prior to the establishment of microcosms, all soils were air-dried at room temperature for several weeks and sieved with a 2-mm mesh sieve. A common inoculum was provided to all laboratories to assure that rhizobia specific to *M. truncatula* were present in all soils.

Abiotic environmental conditions

The set points for environmental conditions were 16 h light (at 300 μmol m^{-2} s^{-1} PAR) and 8 h dark, 20/16°C, 60/70% air RH for day- and night-time, respectively. Different soils (for glasshouses) and treatments with sand patches likely affected water drainage and evapotranspiration. The watering protocol was thus based on dry weight relative to weight at full water holding capacity (WHC). The WHC was estimated based on the weight difference between the dry weight of the containers and the wet weight of the containers 24 h after abundant watering (until water was flowing out of the drainage holes in the bottom of each container). Soil moisture was maintained between 60 and 80% of WHC (i.e. the containers were watered when the soil water dropped below 60% of WHC and water added to reach 80% of WHC) during the first 3 weeks after seedling transplantation and between 50 and 70% of WHC for the rest of the experiment. Microcosms were watered twice a week with estimated WHC values from two microcosms per treatment. To ensure that the patch/heterogeneity treatments did not become a water availability treatment, all containers were weighed and brought to 70 or 80% of WHC

every two weeks. This operation was synchronized with within-block randomization. All 14 experiments were performed between October 2014 and March 2015.

Sampling and analytical procedures

After 80 days, all plants were harvested. Plant shoots were cut at the soil surface, separated by species, and dried at 60ºC for three days. Roots and any remaining litter in the tea bags were washed out of the soil using a 1-mm mesh sieve and dried at 60ºC for three days. Microcosm evapotranspiration rate was measured before the harvesting as the difference in weight changes from 70% of WHC after 48 h. Shoot C%, N%, $\delta^{13}C$, and $\delta^{15}N$ were measured on pooled shoot biomass (including seeds) of *B. distachyon* and analyzed at the Göttingen Centre for Isotope Research and Analysis using a coupled system consisting of an elemental analyzer (NA 1500, Carlo Erba, Milan, Italy) and a gas isotope mass spectrometer (MAT 251, Finnigan, Thermo Electron Corporation, Waltham, Massachusetts, USA).

Data analysis and statistics

All analyses were done using R version 3.2.4[29]. Prior to data analyses, each laboratory was screened individually for outliers. Values that were lower or higher than 1.5 × IQR (interquartile range)[30] within each laboratory, and representing less than 1.7% of the whole dataset, were considered to be outliers due to measurement errors or typos. These values were removed and subsequently treated as missing values. We then assessed whether the impact of the presence of legume (LEG) varied with laboratory (LAB) and the treatment of controlled systematic variability (CSV). This was tested individually for each response variable (Table 1) with a mixed-effects model using the "nlme" package[31]. Following the guidelines suggested by Zuur et al. (2009)[32], we first identified the most appropriate random structure using a restricted maximum likelihood (REML) approach and selected the random structure with the lowest

514 Akaike information criterion (AIC). For this model, CSV and LAB were included as fix factors,
515 experimental block as a random factor, and a "varIdent" weighting function to correct for
516 heteroscedasticity resulting from more heteroscedastic data at the LAB and LEG level (R syntax:
517 "model= lme (response variable ~ LEG*CSV*LAB, random=~1|block, weights=varIdent (form
518 = ~1|LAB*LEG)") (Table 2). As the LAB and SETUP experimental factors were not fully
519 crossed (i.e. laboratories performed the experiment only in one type of setup), the two
520 experimental variables could not be included simultaneously as fixed effects. Therefore, to test
521 for the SETUP effect, we used an additional complementary model including CSV and SETUP
522 as fix effects and laboratory as a random factor (R syntax: "model= lme (response variable ~
523 LEG*CSV*SETUP, random=~1|LAB/block, weights=varIdent (form = ~1|LAB*LEG)")
524 (Supplementary Table 3). To test whether the results were affected by the collinearity among the
525 response variables, the two models also were run on the first (PC1) and second (PC2) principal
526 components the 12 response variables (Fig. 4ab). PCs were estimated using the "FactoMineR"
527 package[33], with missing values replaced using a regularized iterative multiple correspondence
528 analysis[34] in the "missMDA" package[35]. The same methodology was used to compute a second
529 PCA derived from the net legume effect on the 12 response variables (Supplementary Fig. 4cd).
530 To assess how many laboratories produced results that were statistically indistinguishable from
531 one another, we applied Tukey's post-hoc HSD test in the "multcomp" package to lab-specific
532 estimates of PC1 and PC2 (Table 2).
533 To assess how the CSV treatments affected the among- and within-laboratory variability,
534 we used the standard deviation (SD) instead of the coefficient of variation, because the net
535 legume effect contained both positive and negative values. To calculate among- and within-
536 laboratory SDs, we centered and scaled the raw values using the *z*-score normalization [*z*-scored

variable = (raw value–mean)/SD] individually for each of the 12 response variables. Among-laboratory SD was computed from the mean of the laboratory *z*-scores for each response variable, CSV, and SETUP treatments (n = 144; 6 CSV levels × 2 SETUP levels × 12 response variables). Within-laboratory SDs were computed from the values measured in the six replicated microcosms for each CSV and SETUP treatment combination, individually for each response variable, resulting in a dataset with the same structure as for among-laboratory SDs (n = 144; 6 CSV levels × 2 SETUP levels × 12 response variables). Some of the 12 response variables were intrinsically correlated, but most had correlation coefficients < 0.5 (Supplementary Fig. 5) and were therefore treated as independent variables. To analyze and visualize the relationships between the SDs calculated from variables with different units, before the calculation of the among- and within-laboratory SD, the raw values of the 12 response variables were centered and scaled.

The impact of experimental treatments on among- and within-laboratory SD was analyzed using mixed-effect models, following the same procedure described for the individual response variables. The model with the lowest AIC included a random slope for the SETUP within each response variable as well as a "varIdent" weighting function to correct for heteroscedasticity at the variable level (R syntax: "model= lme (SD ~ CSV*SETUP, random=~SETUP|variable, weights=varIdent (form = ~1|variable)) (see also Supplementary Notes). The relationship between within- and among-laboratory SD also was tested with a model with similar random structure but with among-laboratory SD as a dependent variable and within-laboratory SD, CSV, and SETUP as predictors.

Because the treatments containing genotypic CSV increased reproducibility in growth chambers, but slightly increased within-laboratory SD, we also examined the effect of adding

CSV on the statistical power for detecting the net legume effect in each individual laboratory. This analysis was done with the "power.anova.test*"* function in the "base" package. We computed the statistical power of detecting a significant net legume effect (if one had used a one-way ANOVA for the legume treatment) for CTR, GEN_W, GEN_A and ENV_W+GEN_W treatments for each laboratory and response variable. This allowed us to calculate the average statistical power for the aforementioned treatments and how many additional replicates would have been needed to achieve the same statistical power as we had in the CTR.

The data that support the findings of this study are publicly available at https://doi.pangaea.de/10.1594/PANGAEA.880980

Additional References for methods

30. R Development Core Team. R: a language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. (2017).

31. Tukey, J. W. *Exploratory Data Analysis*. (1977).

32. Pinheiro, J., Bates, D., DebRoy, S. & Sarkar, D. NLME: Linear and nonlinear mixed-effects models. *R Packag. version 3.1-122, http//CRAN.R-project.org/package=nlme* 1–336 (2016).

33. Zuur, A. F., Ieno, E. N., Walker, N., Saveliev, A. a & Smith, G. M. *Mixed-effects Models and Extension in Ecology with R*. (2009).

34. Lê, S., Josse, J. & Husson, F. FactoMineR: An R package for multivariate analysis. *J. Stat. Softw.* **25,** 1–18 (2008).

35. Josse, J., Chavent, M., Liquet, B. & Husson, F. Handling missing values with regularized iterative multiple correspondance analysis. *J. Classif.* **29,** 91–116 (2010).

582 36. Josse, J. & Husson, F. missMDA : A package for handling missing values in multivariate
583 data analysis. *J. Stat. Softw.* **70,** 1–31 (2016).

584 **Table 1 | Impact of experimental treatments on response variables.** Mixed-effects model outputs summarizing the F- and P-values
585 (as asterisks) for the impacts of the presence of legumes (LEG), controlled systematic variability (CSV) and laboratory (LAB) on the
586 12 response variables. We also present the impact of experimental treatments on the first and second principal components (PC1 and
587 PC2) of all 12 response variables. The response variables we measured are a typical ensemble of variables measured in plant-soil
588 microcosm experiments (BM = biomass). † symbol indicates response variables measured for the grass *B. distachyon* only, whereas
589 the rest of the variables were measured at the microcosm level, i.e. including the contribution of both the legume and the grass species.
590 Asterisks indicate the significance levels (*** for $P < 0.001$; ** for $P < 0.01$; * for $P < 0.05$; + for $P < 0.1$; ns for $P > 0.1$). DF =
591 numerator degrees of freedom.
592
593

	DF	Shoot BM	Root BM	Seed BM†	Total BM	Shoot/Root	Grass height†	Shoot N%†
LEG	1	4602.95 (***)	1131.65 (***)	2186.64 (***)	690.73 (***)	1137.01 (***)	3.33 (+)	449.87 (***)
CSV	5	15.57 (***)	23.93 (***)	58.01 (***)	1.78 (ns.)	23.98 (***)	23.36 (***)	0.78 (ns.)
LAB	13	1088.67 (***)	182.53 (***)	364.57 (***)	1251.96 (***)	183.42 (***)	317.33 (***)	335.18 (***)
LEG×CSV	5	23.64 (***)	4.48 (***)	33.62 (***)	3.49 (**)	4.51 (***)	2.62 (*)	1.34 (ns)
LEG×LAB	13	235.99 (***)	40.58 (***)	78.17 (***)	116.63 (***)	40.38 (***)	49.89 (***)	14.12 (***)
CSV×LAB	65	6.55 (***)	3.15 (***)	6.93 (***)	7.33 (***)	3.17 (***)	10.16 (***)	1.98 (***)
LEG×LAB×CSV	65	2.22 (***)	1.12 (ns.)	2.70 (***)	1.18 (ns.)	1.12 (ns.)	1.45 (*)	1.71 (***)
		n = 1005	n = 989	n = 997	n = 976	n = 987	n = 1008	n = 1008

	DF	Shoot C%†	Shoot $\delta^{15}N$†	Shoot $\delta^{13}C$†	ET	Litter	PC1	PC2
LEG	1	110.67 (***)	14.43 (***)	26.62 (***)	1269.93 (***)	1.81 (ns.)	1242.53 (***)	988.88 (***)
CSV	5	0.16 (ns.)	8.85 (***)	75.73 (***)	9.37 (***)	1.05 (ns.)	12.87 (***)	22.56 (***)

LAB	13	174.50 (***)	258.30 (***)	888.42 (***)	748.66 (***)	117.34 (***)	920.65 (***)	513.83 (***)
LEG×CSV	5	2.55 (*)	6.48 (***)	5.15 (***)	1.24 (ns.)	1.77 (ns.)	7.08 (***)	11.79 (***)
LEG×LAB	13	11.90 (***)	16.78 (***)	2.52 (**)	172.74 (***)	2.05 (*)	118.12 (***)	28.22 (***)
CSV×LAB	65	1.67 (**)	4.39 (***)	4.97 (***)	21.69 (***)	2.97 (***)	7.22 (***)	2.76 (***)
LEG×LAB×CSV	65	1.33 (*)	1.84 (***)	1.23 (ns.)	1.53 (**)	1.17 (ns.)	0.93 (ns.)	1.65 (**)
		n = 1008	n = 963	n = 973	n = 1002	n = 974	n = 1008	n = 1008

594

595 **Table 2 | Impact of experimental treatments on the number of laboratories that reproduced the**
596 **same finding.** Numbers represent the total number of statistically indistinguishable laboratories based
597 on a Tukey's post-hoc Honest Significant Difference test of the first (PC1) and second (PC2) principal
598 components of the net legume effect of the 12 response variables (see Supplementary Fig. 4cd for the
599 PCA results). For a detailed description of experimental treatments and abbreviations, see Fig. 1.
600

Source	All laboratories (n = 14)		Glasshouses (n = 6)		Growth chambers (n = 8)	
	PC1	PC2	PC1	PC2	PC1	PC2
CTR	7	11	3	5	5	5
ENV_W	10	9	3	3	6	6
ENV_A	8	8	3	4	6	6
GEN_W	8	10	3	3	6	7
GEN_A	11	10	3	3	7	8
ENV_W+GEN_W	11	10	4	3	7	7

601
602

603 # Figure legends
604
605 **Fig. 1 | Experimental design of one block.** Grass monocultures of *Brachypodium distachyon* (green
606 shades) and grass-legume mixtures with the legume *Medicago trunculata* (orange-brown shades) were
607 established in 14 laboratories; shades of green and orange-brown represent three distinct genotypes of
608 *B. distachyon* (Bd21, Bd21-3 and Bd3-1) and *M. truncatula* (L000738, L000530 and L000174). Plants
609 were established in a substrate with equal proportions of sand (black spots) and soil (white), with the
610 sand being either mixed with the soil or concentrated in sand patches to induce environmental
611 controlled systematic variability (CSV). Combinations of three distinct genotypes were used to
612 establish genotypic CSV. Alongside a control (CTR) with no CSV and containing one genotype
613 (L000738 and/or Bd21) in a homogenized substrate (soil-sand mixture), five different types of
614 environmental or genotypic CSV were used as treatments: 1) within-microcosm environmental CSV
615 (ENV_W) achieved by spatially varying soil resource distribution through the introduction of six sand
616 patches into the soil; 2) among-microcosm environmental CSV (ENV_A), which varied the number of
617 sand patches (none, three or six) among replicate microcosms; 3) within-microcosm genotypic CSV
618 (GEN_W) that used three distinct genotypes per species planted in homogenized soil in each microcosm;
619 4) among-microcosm genotypic CSV (GEN_A) that varied the number of genotypes (one, two or three)
620 planted in homogenized soil among replicate microcosms; and 5) both genotypic and environmental
621 CSV (GEN_W+ENV_W) within microcosms that used six sand patches and three plant genotypes per
622 species in each microcosm. The "× 3" indicates that the same genotypic and sand composition was
623 repeated in three microcosms per block. The spatial arrangement of the microcosms in each block was
624 re-randomized every two weeks. The blocks represent two distinct chambers in growth chamber
625 setups, whereas in glasshouse setups the blocks represent two distinct growth benches in the same
626 glasshouse**.**
627

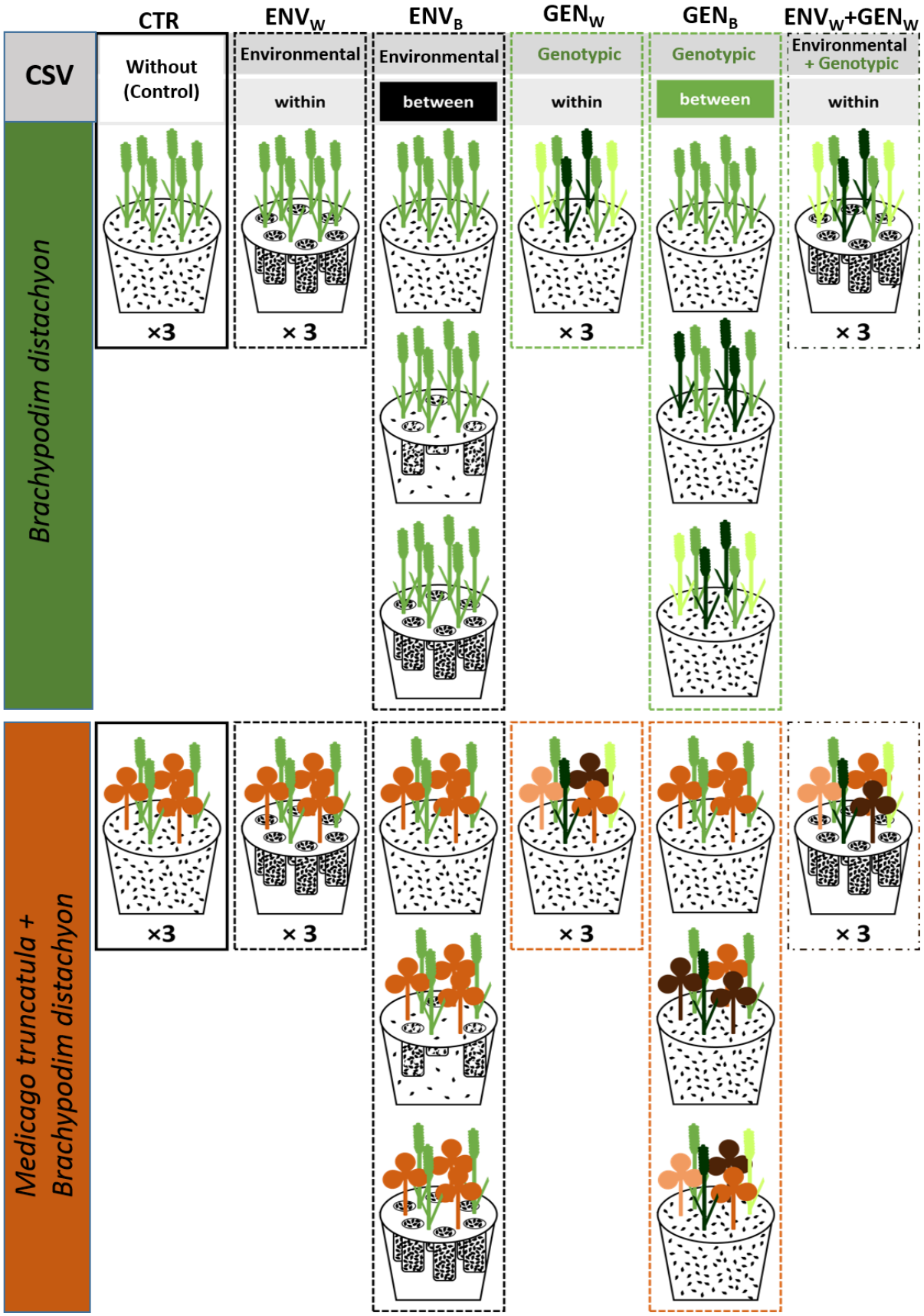
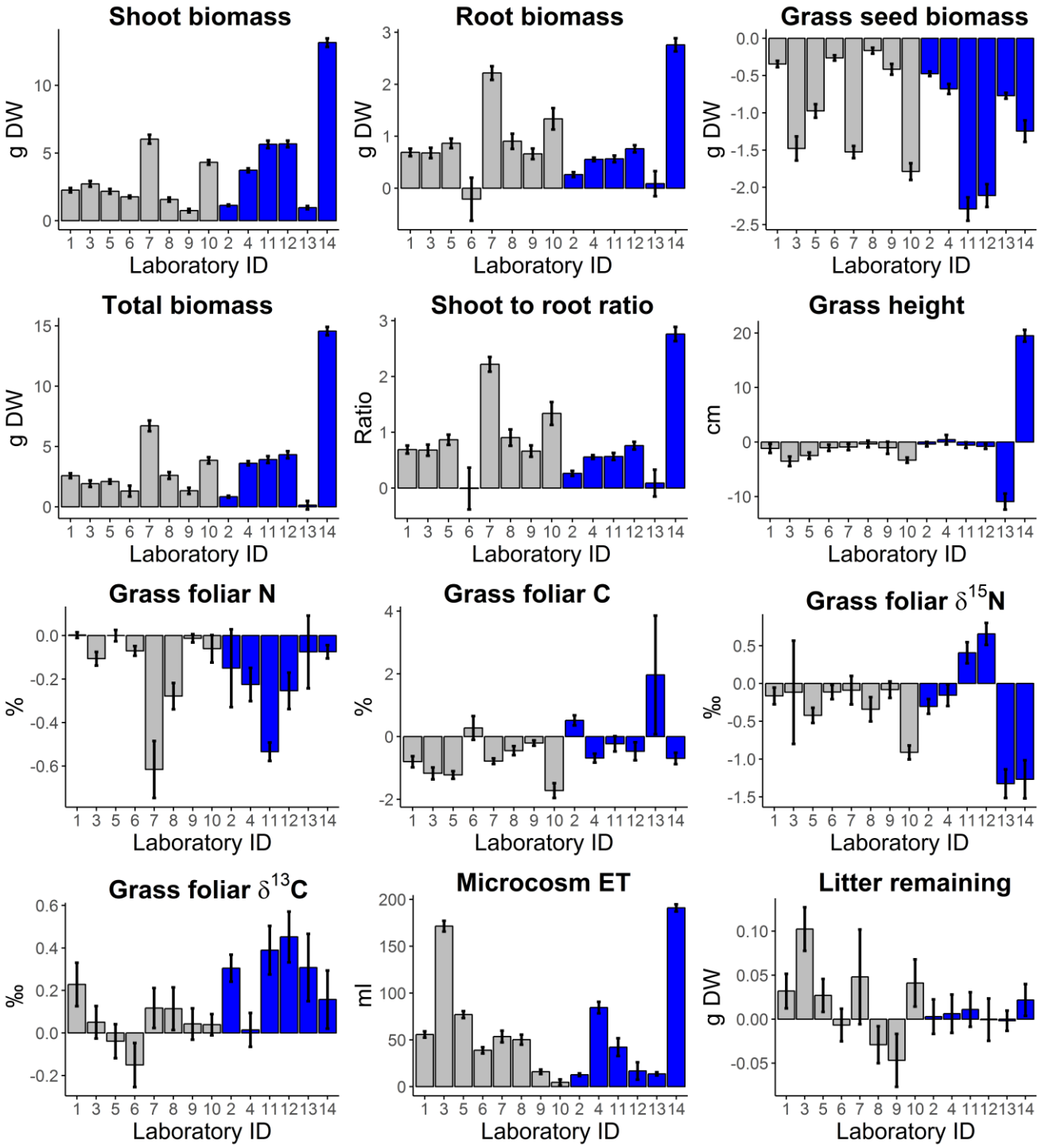

CTR
ENV$_W$
ENV$_B$
GEN$_W$
GEN$_B$
ENV$_W$+GEN$_W$
CSV
Without (Control)
Environmental
within
Environmental
between
Genotypic
within
Genotypic
between
Environmental + Genotypic
within
Brachypodim distachyon
×3
× 3
× 3
× 3
× 3
× 3
Medicago truncatula + Brachypodim distachyon
×3
× 3
× 3
× 3
× 3
× 3

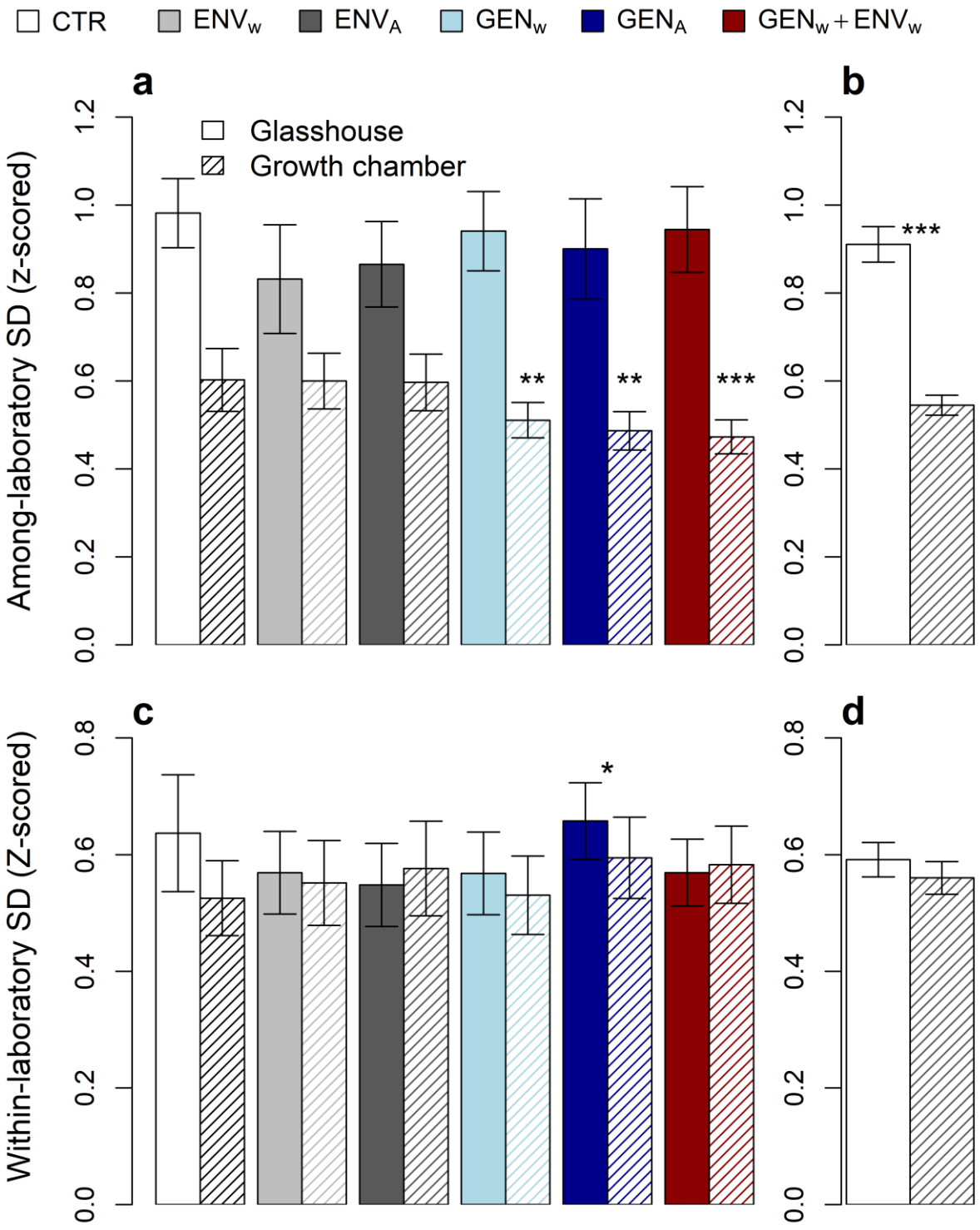
Fig. 2 | Net legume effect for the 12 response variables in 14 laboratories as affected by laboratory and SETUP (growth chamber vs. glasshouse) treatment. The grey and blue bars represent laboratories that used growth chamber and glasshouse set-ups, respectively. Bars show means by laboratory obtained by averaging over all CSV treatments, with error bars indicating ± 1 s.e.m. (n = 72 microcosms per laboratory).


Shoot biomass
g DW
Laboratory ID
Root biomass
g DW
Laboratory ID
Grass seed biomass
g DW
Laboratory ID
Total biomass
g DW
Laboratory ID
Shoot to root ratio
Ratio
Laboratory ID
Grass height
cm
Laboratory ID
Grass foliar N
%
Laboratory ID
Grass foliar C
%
Laboratory ID
Grass foliar δ15N
‰
Laboratory ID
Grass foliar δ13C
‰
Laboratory ID
Microcosm ET
ml
Laboratory ID
Litter remaining
g DW
Laboratory ID
1 3 5 6 7 8 9 10 2 4 11 12 13 14

636

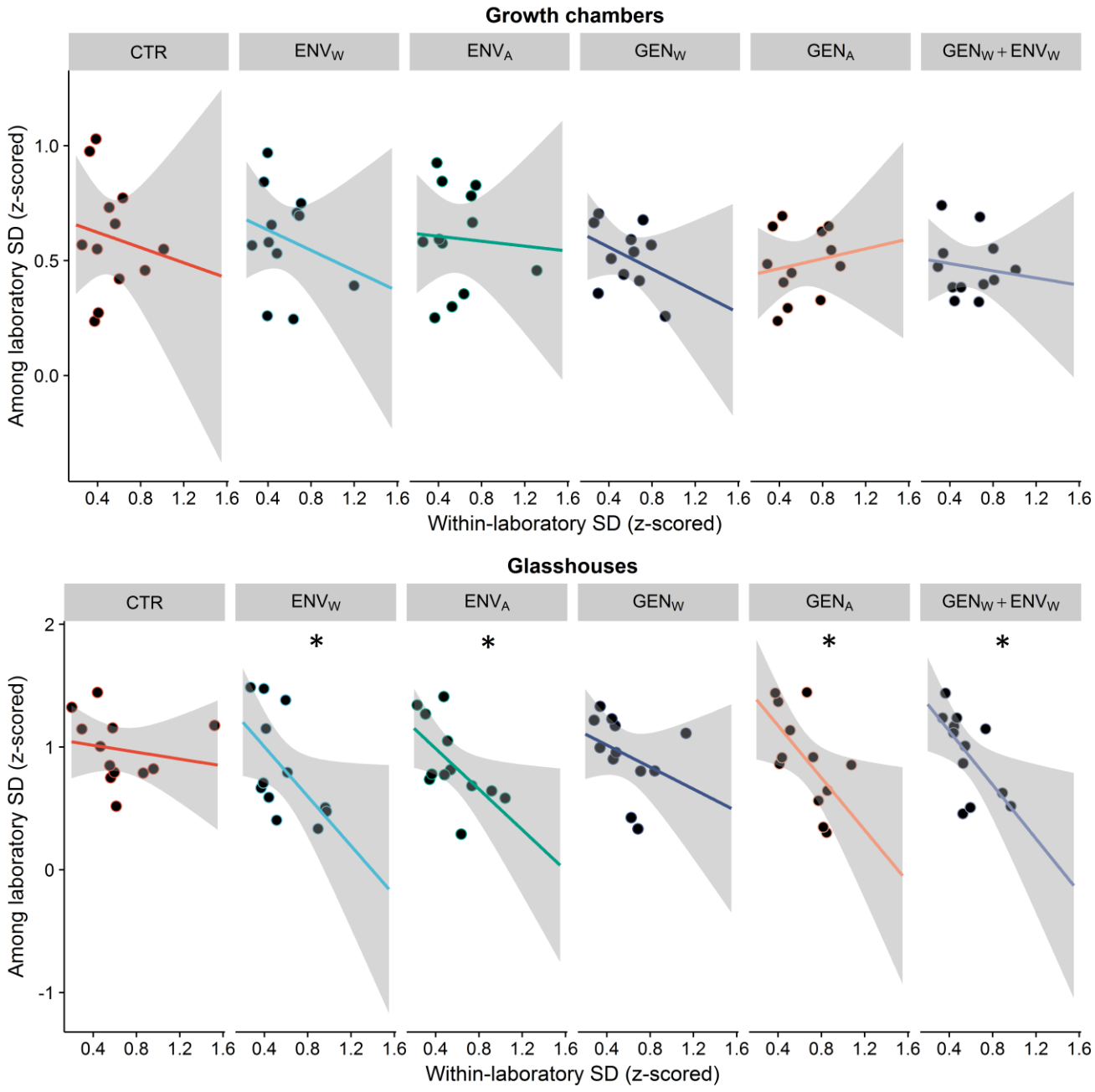
637

Fig. 3 | Among- and within-laboratory standard deviation (SD) of the net legume effect as affected by experimental treatments. Among-laboratory SD as affected by CSV and SETUP (**a**) and SETUP only (**b**). Within-laboratory SD as affected by CSV and SETUP (**c**) and SETUP only (**d**). Lower among-laboratory SD indicates enhanced reproducibility. Solid-filled bars and striped bars represent glasshouse (n = 6) and growth chamber setups (n = 8), respectively. Asterisks represent *P*-values (*** for $P < 0.001$, ** for $P < 0.01$, * for $P < 0.05$) indicating significantly different fitted coefficients according to the mixed-effects models (see Supplementary Notes for full model outputs); in (**c**) the star indicates the significant difference between GEN_A and CTR, irrespective of the type of SETUP. For a detailed description of experimental treatments and abbreviations see Fig. 1.


CTR
ENVw
ENVA
GENw
GENA
GENw + ENVw
a
b
c
d
Glasshouse
Growth chamber
Among-laboratory SD (z-scored)
Within-laboratory SD (Z-scored)
0.0
0.2
0.4
0.6
0.8
1.0
1.2
**
**
***
***
*

648

649



650

651 **Fig. 4 | Relationship between within-laboratory SD and among-laboratory SD of the net legume**
652 **effect as affected by experimental treatments.** The figure illustrates the significant within-laboratory
653 SD×SETUP×CSV three-way interaction ($F_{5,109}$ = 2.4, P < 0.040) affecting among-laboratory SD
654 (Supplementary Note). This interaction is the result of a more negative relationship between within-
655 and among-laboratory SD in glasshouses relative to growth chambers, but with different slopes for the
656 different CSV treatments. Points represent the 12 response variables. Asterisks represent *P* values <

657 0.05 for the individual linear regressions. Note the different scale for the y-axis between growth
658 chambers and glasshouses. For a detailed description of experimental treatments and abbreviations see